基于多元数据的城市区域可达性评估模型 *

单晓晨，曲海成†，刘万军

(辽宁工程技术大学 软件学院, 辽宁 葫芦岛 125000)

摘　要：城市区域可达性评估一直以来是智能交通领域备受关注的热点问题。传统的区域可达性评估模型一般只支持GIS、GPS 等单一数据作为可达性的评估依据，无法避免因外界因素的影响对区域可达性造成的评估不准确问题。针对此问题，以出租车 GPS 行车数据、时段、天气等多维数据作为区域可达性的评估依据，构建了一种支持多元数据的城市区域可达性评估模型，在此基础上设计了基于多维 OD 矩阵的多元数据区域可达率计算方法，并将可达率作为区域可达性量化标准以达到提高可达性评估准确性的目的。此外，针对因传统 GPS 数据清洗方法过于粗糙而导致的有效信息遗漏、数据矫正不准确问题，利用基于统计学理论的序列数据清洗方法，运用出租车 GPS 数据的速度与加速度信息纠正潜在的误差数据以提高 GPS 数据的清洗效果。实验证明，利用提出的多元数据城市区域可达性评估模型可达性评价的准确率提高 9.1%-37.8，其中计算的区域可达率的准确性较传统方法提高 12.6%-35.5%，平均旅行时间的准确率提高 18.5%-31.6%。

关键词：GPS；可达性；可达率；轮廓测量法；OD 矩阵

中图分类号：TP301.4

Evaluation model of urban area accessibility based on multivariate data

Shan Xiaochen, Qu Haicheng†, Liu Wanjun

(*School of Software Liaoning Technical University*, *Huludao Liaoning* 125105, *China*)

Abstract: The assessment of urban accessibility has always been a hot topic of concern in the research field of smart transportation. The traditional regional accessibility assessment model generally only supports the single dimension data of GIS or GPS as the basic data for the assessment of accessibility, so it is impossible to avoid the problem of inaccurate assessment of regional accessibility due to the influence of external factors. Aiming at this problem, this paper constructs a city area accessibility evaluation model to support multivariate data using the multidimensional data such as GPS vehicle traffic data, time and weather as the basis of regional accessibility. On this basis, the calculation model of the region accessibility ratio based on multidimensional OD matrix is designed in this work which is used as the quantitative methods of regional accessibility to achieve the purpose of improving the accuracy of accessibility assessment. In addition, to solve the problem of traditional GPS data cleaning method, such as effective information missing and inaccurate data correction, which is caused by its over-roughness, the serial data cleaning method based on the statistical theory is applied in this model. The speed and acceleration information of the Taxi GPS data is considered in this data cleaning method to correct the potential error and to improve the GPS data cleaning effect. Experiment result shows that the accuracy of the regional accessibility calculated by using the multivariate data urban area accessibility assessment model proposed in this paper is 9.1% -37.8 higher than that of the traditional methods, and the accuracy of the regional accessibility assessment ratio and travel time are increased by 12.6% -35.5% and 18.5% -31.6% respectively.

Key Words: GPS; accessibility; ration of accessibility; multivariate data; OD matrix

0　引言

随着人口的不断增长，城市区域不断扩大使人们对出行效率格外关注。很多城市已经由单中心发展成多中心，但是道路交通网的发展并没有跟上城市现代化的发展。这就会造成部分地区出行困难，而解决出行困难的关键就是发现这些出行不便

基金项目：国家科技支撑计划资助项目（2013BAH12F00，MK2013008）；辽宁省教育厅科学技术研究一般项目（L2015216）；第六批生产技术问题创新研究基金资助项目（20160092T）

作者简介：单晓晨（1994-），女，辽宁大连人，硕士，主要研究方向为大数据分析与智能交通；曲海成（1981-），男（通信作者），副教授，博士，主要研究方向为大数据分析（haichengqu@hit.edu.cn）；刘万军（1959-），男，教授，博导，主要研究方向为大数据图像与视觉信息计算.

利的地区也就是可达性差的地区并分析造成可达性差的原因。通过改善个别地区的可达性进而提高整个城市的交通通畅水平。并且区域可达性的评估可以为交通管理部门解决交通拥塞、道路建设规划等问题提供决策依据。

城市区域可达性的评估通常先将城市平均分成若干区域，由获得的旅行信息聚类出热点（参与旅行频率较高的地区），包含热点的区域为学习区域。评估工作主要是对学习区域内的旅行信息进行挖掘，计算区域到区域的平均旅行时间，最后与规定的阈值相比较，结合量化策略对可达性进行量化。

目前对于可达性的量化策略主要分为两种。传统基线模型对 GIS[1]数据进行分析所产生的可达性评估主要是从一个区域到另一区域的平均旅行时间的角度来量化。在获得起始区域到所有目的区域的平均旅行时间后再取平均旅行时间的平均值从而评估该区域的可达性，但是这种基线法在求区域间平均旅行时间时是用区域间的直线距离除以平均速度而得到的时间。因此其计算的时间会有误差，这只适用于旅行路线接近于直线的情况。但在实际旅行中，大多数行车路线都长于起止点间直线距离，因此传统基线法对区域可达性的评估存在着必然的误差。随着卫星定位技术的普及，各大城市的出租车均已安装 GPS 系统用于指挥调度与数据分析。出租车 GPS 数据可以反映出一个城市的旅行分布，出租车司机在载客过程中往往会凭借经验选择出最省时的路线。这可以避免路线选择对可达性造成的影响。为了降低 GIS 数据只能产生直线距离而对可达性造成的误差，对于可达性的研究逐渐采用面向 GPS 数据的分析方法。对 GPS 数据进行分析是从可达热点的数量这一角度来量化，以可达热点的个数来表示可达性，最为代表的算法为轮廓测量法[2]。该方法普遍认为当前研究的学习区域到另一个学习区域的平均旅行时间只要大于城市平均旅行时间，则认为涉及到的目标区域的所有热点对于当前的起始学习区域都是不可达的，反之所有热点全是可达的。虽然这种方法可以估计出可达性明显较好或者明显较差的区域，但是每个区域与其他区域连通的热点个数与旅行时间都各不相同，显然这种方法对于旅行分布不均以及可达性好坏不明显的区域的评估效果较差。此外，现有的城市区域可达性的研究主要采用一元数据如 GIS 数据或 GPS 数据进行建模分析。其构建的模型只能支持一维数据处理，因此无法避免因时间、天气等外部因素的影响对可达性评估造成的误差。

为了缩小因可达性的量化方法而产生的误差，本文利用区域可达率作为可达性的量化标准，即由该区域出发的所有旅行中，旅行时间在阈值范围内的旅行次数占总次数的比例。以区域可达率来量化区域可达性不仅可以提高传统轮廓测量法的准确性，而且考虑到区域的旅行分布特点可以避免轮廓测量法对旅行分布不均的区域产生错误估计。而且从 GPS 数据中抽取出的旅行不需考虑旅行路线，只计算旅行起点与终点的时间差便可计算出准确的旅行时间，克服了传统基线法依赖平均旅行时间却无法准确求出平均旅行时间的弊端。在此基础上本文以出租车 GPS 行车数据、时段、天气等多维数据作为区域可达性的评估依据，构建了一种支持多元数据的城市区域可达性评估模型，又针对多维数据的可达性量化问题设计了基于多维 OD 矩阵(Origin Destination 矩阵)的多元数据区域可达率计算方法。通过建立多维 OD 矩阵，结合出租车 GPS 数据从 GPS 数据中抽出完整旅行并从旅行信息的角度出发计算区域可达率。此外，在数据清洗方面采用一种基于统计学的有序数据的清洗方法[3]对部分有误的 GPS 序列进行纠正。时段、天气等外部因素的加入可以进行多种旅行条件下的区域可达性分析，通过控制变量的方法不仅可以测出区域的普遍可达性，还可以分析出如‘潮汐交通’等特殊条件下区域可达性的变化，以及产生变化的原因。这可以为道路建设，交通指挥提供诸多关键信息，进而提高人们的生活工作效率。

1 相关工作及研究

可达性最初被定义为交通网络中每一区域可以与其他区域进行通信的机会大小，后来 shen[4] 等人从城市空间的角度出发认为可达性与市民的社会经济活动以及这些活动的地理关系密切相关，区域可达性恰可以衡量这些地理关系的关联程度。随着可达性概念被不断的完善，目前可达性可以被普遍认为是在某种运输系统或某些运输方式下，使个体达到目的地的容易程度[5]。区域可达性被认为是衡量城市效率的一种方法，被广泛应用于交通领域。

对区域可达性的研究可以使用地理信息系统对可达性进行预测[6~8],地理信息系统是以搜索最短路径为前提，并对产生的最短时间进行加权。Novak 等人[9]使用个人信息数据对不同时段的可达性进行计算。这可以更加真实地*反映*在不同时间段内旅行路线与旅行时间的关系，但需要大量的个人信息数据才能计算出人群普遍的旅行分布，才能使得评估结果具有代表性。

如今出租车 GPS 已被普及，由于其易于解释区域间的沟通，而且可用的数据量大，出租车 GPS 数据已被广泛应用于城市规划[10]、地理学研究[11]、旅游需求的建模[12]和旅行时间的估计[13,14] 等领域。Laha[15]等人经研究表明，对出租车 GPS 数据经过计算及知识挖掘后，可得到路段旅程时间、路段平均速度和道路拥塞程度等信息，还可以获知司机选则的路线的倾向以及乘客乘降的密集地点等信息，进而反映城市交通流的信息，既帮助乘客了解出行信息，又帮助司机优化导航路线。除此之外出租车 GPS 数据还被用于设计夜班巴士线等交通热线[16,17]。通过分析单向或双向路线流，设计夜班巴士总线。

基于出租车 GPS 数据的分析方法通常被认为更加适合于预测运输计划的可实现程度以及评估区域可达性。基于出租车 GPS 的方法核心是对旅行时间的挖掘与学习。对于可达性的量化方式可以用区域间的平均旅行时间作为衡量标准。也可以以速度恒定，距离加权的方法计算旅行时间及可达的热点数来量化可达性。这些方法只能支持 GPS 坐标单一信息的分析，其结果利用统计学方法计算估计出来的而非通过数据挖掘方法计算

得出。因此无法避免因时间段、天气等外部因素的影响而对计算结果造成的误差。此外，由于分析模型的所用到的数据规模庞大，原始 GPS 数据存在着较多垃圾数据，如果不加以处理，这些垃圾数据会导致区域可达性的评估误差较大，因此对 GPS 数据进行清洗是至关重要的。在数据量大的前提下，大多数做法如 Cui J X 等人[2]对错误数据采取直接删除的方法，这种方法虽然可以过滤掉大量的错误信息，但是会使得 GPS 数据的连续性丢失，同时由于过滤粒度的设置不当也会产生有用信息的误删。

2　区域可达性评估模型

本文以出租车 GPS 行车数据、时段、天气等多维数据作为区域可达性的评估依据，构建了支持多元数据的城市区域可达性评估模型，在此基础上设计了基于多维 OD 矩阵的多元数据区域可达率计算方法，并利用区域可达率作为区域可达性的量化标准来衡量城市区域的可达程度。可达率即为在规定的旅行时间内区域 P_i 到其余区域的旅行中可以完成的旅行的比例。区域可达率可以更客观真实地反映区域可达性的好坏。区域可达性评估模型整体架构如图 1 所示。

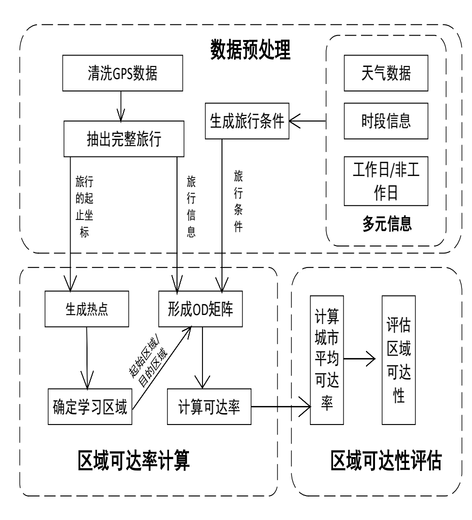


图 1　可达率预测模型整体架构

该模型主要分为以下部分：a) 数据预处理，这一部分包括清洗 GPS 序列、提炼出完整的旅行以及整合多元数据生成旅行条件;b）区域可达率计算，这一部分包括聚类生成热点、确定学习区域、产生 OD 矩阵、计算可达率结果。对没有规律可言的所有旅行的起始点和终点采用 DBSCAN 密度聚类法[18]形成热点。其目的是把看似不相关的旅行联系起来，使这些旅行在起始位置和终点位置上具有相同属性，则这类旅行便可反映起始区域至终点区域的道路可达性。模型将至少包含一个热点的区域作为学习区域。通过之前抽取出的完整的旅行可获得旅行的起止点，同时可以获取到旅行起止点所属的热点进而确定旅行发生的起止学习区域，再结合其他旅行信息建立多维 OD 矩阵。最后利用 OD 矩阵从旅行的角度出发计算区域可达率;c）评估区域可达性，这一部分主要包括计算各区域可达率的平均值，并以此作为阈值评估区域可达性。

2.1　出租车 GPS 数据的筛选与清洗

出租车 GPS 数据十分庞大，然而其中不乏一些错误的数据会干扰区域可达率的预测结果，这些错误数据将被清除或修改。本文从以下三个方面对数据进行筛选与清理：

a）由出租车 GPS 数据可知出租车的地理位置即坐标。当此条 GPS 数据有错误时坐标记为 0 则证明该 GPS 点不能被使用也不能被纠正所以将这类 GPS 点直接过滤掉。

b）由 GPS 数据可知相邻 GPS 点 $G_m(C_m,T_m,S_m)$ 和 $G_{m+1}(C_{m+1},T_{m+1},S_{m+1})$间的速度 V_m，即

$$V_m=\frac{\sqrt{(x_{m+1}-x_m)^2+(y_{m+1}-y_m)^2}}{T_{m+1}-T_m} \tag{1}$$

当 V_m 超出阈值（V_{max}）时，则此条数据有误，应该被纠正。其中 x_m，x_{m+1}，y_{m+1}，y_m 为 GPS 点 $G_m(C_m,T_m,S_m)$和 $G_{m+1}(C_{m+1},T_{m+1},S_{m+1})$中 C_m 和 C_{m+1} 值即坐标。$T_{m+1}-T_m$ 为两 GPS 间的时间差。对于两点间的速度错误可以由统计纠正法预测出正常的速度，然后再根据预测出的速度对坐标信息进行纠正。

由相邻的 GPS 点还可求相邻时间序列间的加速度 a_m，即

$$a_m=\frac{V_{m+1}-V_m}{T_{m+1}-T_m} \tag{2}$$

当加速度超出阈值（M_a）时，此数据被认为有误。但这类错误也可以被统计纠正法修正。

c)因为不是所有的司机都会为乘客选择最近最优路线，有些司机为了获利会选择较远的旅行路线，这将直接增加乘客旅途的时间。为避免因绕道现象和同一乘客乘同一出租车先后到达不同地方而造成旅行时间过长导致的可达率较低的假象。本文以 $REDUN_D_{jour}$，即

$$REDUN_D_{jour}=\frac{DIS_real}{DIS_ec} \tag{3}$$

作为过滤条件,超出阈值（3.5）则认为此次旅行不能真实反映交通情况，应过滤掉整条旅行。

其中 DIS_ec 为旅行的直线距离，即

$$DIS_ec=\sqrt{(x_n-x_1)^2+(y_n-y_1)^2} \tag{4}$$

x_n，y_n 为旅行中的最后一个 GPS 点的位置坐标。x_1，y_1 为旅行中的第一个 GPS 点的位置坐标。由这两点的坐标便可计算出旅行中出发点到目的地的直线距离 DIS_ec。DIS_real 由旅行中所有相邻 GPS 点的直线距离累加得出,为实际旅行所用距离即

$$DIS_real=\sum_{k=1}^{n}\sqrt{(x_{k+1}-x_k)^2+(y_{k+1}-y_k)^2} \tag{5}$$

2.2 多元旅行数据结构

每一辆出租车每一天都会产生大量 GPS 位置坐标点，这些点中包含了很多条旅行。

定义 1 旅行 trip 指每位乘坐出租车的旅客所经历的旅途。旅行包括起止位置坐标、旅行总时长、旅行发生的日期、旅行所经历的路程。

这类旅行可以看成是由一系列满足时间序列的 GPS 点组成的。本文定义 GPS 点的结构为 G(C,T,S), GPS 轨迹可以被表示为 $G_1(C_1,T_1,S_1)$ ⋯$G_m(C_m,T_m,S_m)$ ⋯$G_n(C_n,T_n,S_n)$。其中 C 为 GPS 位置坐标，T 为时间， S 为 0 代表出租车空载，为 1 代表车内有乘客，出租车已被占用。出租车每隔一段时间产生一个 GPS 数据，而一次旅行共产生 n 次 GPS 数据，所以下标取值为 1 到 n。为了提取出每一条旅行中有价值的信息，本文以 OD 矩阵为原型构建一种数据结构来表达旅行信息，即 4 维 OD 矩阵。4 维 OD 矩阵各维度示意图如图 2 所示。第一维和第二维分别为起始区域和目的区域，将整个城市平均分为 W×Z 个区域用 P_i 表示起始区域（i 属于 1……W×Z），P_j 表示目的区域（j 属于 1……W×Z）。第三维旅行条件，因为同一地区的道路可达率在工作日与非工作日里是不同的，在同一天的不同时间内也是不同的，且天气因素也会影响交通区域可达率，因此矩阵的第三维包含了多元旅行信息如 DAY_T=0，代表非工作日；DAY_T=1，代表工作日。TIME_T 来表示一天内的不同时间段（如以下五个时间 7:00-9:00, 9:00-12:00, 12:00-16:00, 16:00-19:00, 19:00-次日 7:00 ）; WEATHER=0 代表有雨; WEATHER=1 代表天晴。上述三个条件 DAY_T, TIME_T, WEATHER 的不同取值可以组成 20 种不同旅行条件（例如晴天工作日的早上，雨天休息日的晚上等）type。type 的种类及意义见表 1。第四维是具体的旅行信息包含旅行时间、实际旅行距离、旅行日期等。$TIME_{jour}$ 来表示此次旅途所用的总时间，即

$$TIME_{jour} = T_n - T_1 \tag{6}$$

其中: T_1, T_n 为一次旅行中记录的 GPS 信息的起始点时间和终止点时间。

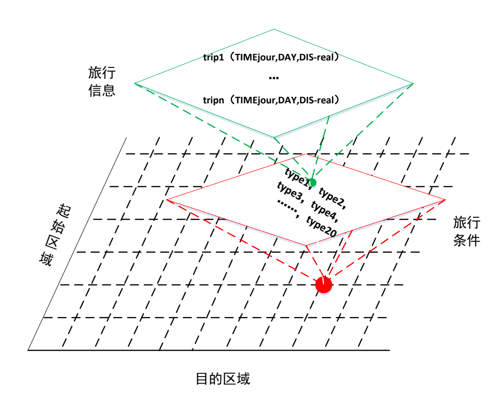


图 2 4 维 OD 矩阵示意图

DAY 代表此次旅行的日期，取值范围由所掌握的实际 GPS 数据量而定，在此数据结构中起到主键的作用即确定每次旅途的唯一性。由此 OD 矩阵可写作（P_i, P_j, type, trip。）

表 1 变量 type 的种类及意义

type	意义	DAY_T/TIME_T/WEATHER
WMF	工作日/7:00-9:00/晴	1/1/1
WNF	工作日/9:00-12:00/晴	1/2/1
WAF	工作日/12:00-16:00/晴	1/3/1
WEF	工作日/16:00-19:00/晴	1/4/1
WBF	工作日/19:00-次日 7:00/晴	1/5/1
VMF	非工作日/7:00-9:00/晴	0/1/1
VNF	非工作日/9:00-12:00/晴	0/2/1
VAF	非工作日/12:00-16:00/晴	0/3/1
VEF	非工作日/16:00-19:00/晴	0/4/1
VBF	非工作日/19:00-次日 7:00/晴	0/5/1
WMR	工作日/7:00-9:00/雨	1/1/0
WNR	工作日/9:00-12:00/雨	1/2/0
WAR	工作日/12:00-16:00/雨	1/3/0
WER	工作日/16:00-19:00/雨	1/4/0
WBR	工作日/19:00-次日 7:00/雨	1/5/0
VMR	非工作日/7:00-9:00/雨	0/1/0
VNR	非工作日/9:00-12:00/雨	0/2/0
VAR	非工作日/12:00-16:00/雨	0/3/0
VER	非工作日/16:00-19:00/雨	0/4/0
VBR	非工作日/19:00-次日 7:00/雨	0/5/0

图 2 中底层平行四边形为 OD 矩阵第一维信息起始区域和第二维信息目的区域，红色菱形框为第三维信息旅行条件。绿色菱形框表示第四维信息旅行信息包含旅行时间，旅行的路程，以及旅行日期。由 GPS 数据抽取出旅行信息的具体算法见算法 1（伪码）。

算法 1 旅行信息获取算法

输入：GPS 数据 GPS[i]

输出：各旅行条件下的旅行信息 on[m], off[m], T1[m], T2[m], $TIME_{jour}$[m]

```
m=0;                    //旅行计数器
for(i=0;GPS[i]!=null;i++)
{ if(GPS[i+1].S==1∧GPS[i].S==0)
//开始载客的 GPS 点
{on[m]=GPS[i+1].C;
//记录旅行开始的位置坐标
T1[m]= GPS[i+1].T;   //记录旅行开始的时间

}
else if(GPS[i].S==1∧GPS[i+1].S==0)
//开始卸客的 GPS 点

{ off[m]=GPS[i].C;// 记录旅行结束的位置坐标
T2[m]= GPS[i].T;    // 记录旅行结束的时间
```

```
TIMEjour[m]=T2[m]-T1[m];        //旅行时间
m++;
}
}
```

算法 1 以一台出租车为例首先通过 GPS 的状态位 s 判断出租车开始载客的点，进而得出旅行开始的时间、位置、天气、以及是否为工作日等信息。再继续搜索后续 GPS 点的状态位 s 得出旅行结束的时间、位置以及旅行时段。当出现前一 GPS 点的状态位为 0 而后一状态位为 1 时可确认后一 GPS 点为新旅行的起始点。当出现前一 GPS 点的状态位为 1 而后一状态位为 0 时可确认前一 GPS 点的为新旅行的终止点。

2.3 确定学习区域

为了方便对城市区域可达性的研究，首先将城市分为 W×Z 个区域。然后再从众多 GPS 点中抽取出完整的旅行，每一条旅行的起始点与终点杂乱无章的分散到各区域中。但也不是每个区域都包含旅行的起点或终点的如河流、建筑、公园等区域不可能包含旅行信息，因此不研究这类对区域可达性没有价值的区域。还存在一些区域虽然包含旅行，但旅行次数过少，那么这样的区域也是没有研究价值的。于是应当从所有的区域中确认出有研究价值的学习区域。一个区域包含旅行的起点或终点足够多则该区域是有价值的可被选为学习区域。在乘降点数量足够多的情况下相邻的所有乘降点可以认为是一个热点。各区域中只要包含一个热点，就可以被选作学习区域。本文以聚类的方法产生热点，热点为发生旅行较密集的地点。旅途实际是由起始坐标和终点坐标所确立，然而这两个坐标可以出现在任意区域内。当某一小范围内旅行起始点和终止点次数达到一定阈值时则符合聚类条件，这样此范围就会被聚类为热点。通过判断之前划分的区域是否包含热点来确定学习区域。确认学习区域的具体算法见算法 2（伪码）。

算法 2 确立学习区域

输入：所有旅行的起止点 on[m],off[m]

输出：学习区域 learn[learn_A]

```
TD=and(on,off);//将旅行的起止点并为一个集合
hot[i]=DBSCAN(TD,E,MINp);
/*由 DBSCAN 密度聚类发求热点，其中 E 为聚类时的扫描半径;MINp 是密度阈值*/
learn_A=0;    //计数学习区域个数
p[region*x+y]=0;
//将所有区域包含热点的初始值设置为 0
for(i=0;hot[i]!=null;i++)
{ x=floor(hot[i].x);
y=floor(hot[i].y);
p[region*x+y]++;
//发现包含热点的区域，并记录目/前包含热点的个数
}
for(j=0;j< regionx* region;j++)
{if(p[j]>0)
{learn[learn_A]=j;
//重新编号学习区域
learn_A ++;
}
}
```

算法 2 首先假设已被划分好的 W×Z 个区域中和区域包含 0 个热点，再以 DBSCAN 这一聚类方法利用区域间发生旅行的起止坐标聚类出热点。比较向下取整后的热点坐标与区域的坐标可得出该热点属于哪一个区域，并使区域所包含的热点数增加。最后查找所有热点数不为 0 的区域将其确定为学习区域，并为确定的学习区域标号。

2.4 建立可达率计算模型

区域道路可达率是在规定的旅行时间范围内，区域 P_i 到达其余区域的完成程度，也就是该区域在规定的时间内可完成的旅行数量占由该区域始发的所有旅行的比重，可完成的次数越多该区域的可达率也就越高，可达性也就越好。可达率的计算主要分为两部分，4 维 OD 矩阵的构建以及利用 4 维 OD 矩阵计算区域可达率。

利用选择出的学习区域生成 OD 矩阵，形成由学习区域组成的非稀疏矩阵，以方便对旅行进行分析与挖掘并提高运算效率。OD 矩阵的生成见算法 3（伪码）。

算法 3 建立 OD 矩阵

输入：旅行信息 trip[m]

输出：OD 矩阵 OD[i][j][TY][KTY]

```
OD[Learn_A][ Learn_A][20][];
for(i=0,i< Learn_A,i++)
{for(m=0;m<length(on);m++)
{if(floor(on[m].x)*regionx+floor(on[m].y)=learn[i])
//查找由学习区域 i 始发的旅行
{for(n=0;n<Learn_A;n++)
if(learn[n]=floor(off[m].x)*regionx+floor(off[m].y))
//查找当前选中旅行的目的区域所对应的学习区域
j=n;            //n 为目的学习区域
}
if(type.weather[m]=w and type.workday[m]=d and period[m]= iod)//w、d、iod 代表三种旅行条件的组
//合每一种组合对应一个 TY 值
OD[i][j][TY][KTY]=trip[m];/*TY 取值为 0-19，将对应的 trip 记录到对应旅行条件下的 KTY 中*/
KTY ++;
}
}
```

算法 3 的主要功能是抽取出的旅行合理的存储起来，方便以后的计算。由算法 1 得出的旅行信息和旅行条件作为 OD 矩阵的 3、4 维信息正确的存储到 OD 矩阵中。矩阵的 1、2 为信

息是旅行的起止区域，这由算法 2 确定

传统的轮廓计算方法就是计算区域可达热点数并以此评估区域可达性。传统的轮廓预测法先计算区域 P_i 在特定的旅行条件 t 下到达各个区域的热点总数计为 $Ac_{_it}$，即

$$Ac_{_it}=\sum_{j=1,\overline{TIME_{ij}}\le TM_{_max}}^{n\wedge type(t)}(a_{tj}) \tag{7}$$

其中: a_{tj} 为区域 P_j 中在满足旅行条件 t 时所涉及的目的区域的热点数。type(t)为旅行条件的类别，$\overline{TIME_{ij}}$ 为区域 P_i 到 P_j 所有旅行的平均时间，即

$$\overline{TIME_{ij}}=\frac{\sum_{n=1}^{count_{ODij}}TIMEjour}{count_{ODij}} \tag{8}$$

$count_{ODij}$ 为区域 i 至区域 j 的旅行次数，可从 OD 矩阵中获取。$TM_{_max}$ 为规定的旅行时间范围的上限。

再计算区域 P_i 在所有旅行条件下到达各个区域的热点总数计为 $Ac_{_i}$，即

$$Ac_{_i}=\sum_{t=1}^{20}Ac_{_it} \tag{9}$$

并以 $Ac_{_i}$ 作为区域 P_i 的可达热点数最终以 $Ac_{_i}$ 的值来评估区域可达性。

式(9)中的 $Ac_{_i}$ 是传统轮廓测量法对区域 P_i 的可达性量化，值越高则表明可达性较好，道路畅通。然而实际上 $\overline{TIME_{ij}}$ 小于 $TM_{_max}$ 并不代表 P_i 到 P_j 的所有旅行中符合条件 t 的旅行时间都小于 $TM_{_max}$，也不意味着所有涉及到的热点 a_{tj} 都可到达，但式(7)却对 $Ac_{_it}$ 累加涉及到的所有热点 a_{tj}，也就是这种情况会高估 $Ac_{_it}$ 的值近而高估 $Ac_{_i}$。同样的，$\overline{TIME_{i}}$ 大于 $TM_{_max}$ 也不代表 P_i 到 P_j 的所有旅行中符合条件 t 的旅行时间都大于 $TM_{_max}$，也不意味着所有涉及到的 a_{tj} 都不可到达，但式(7)却不对 $Ac_{_it}$ 做任何累加，这种情况会低估 $Ac_{_it}$ 的值，进而将误差传递至 $Ac_{_i}$。基于可达热点数的区域可达性判定方法的主要标准为可达热点数的多少，也就是可达热点数越多可达性越好，反之可达性就越差。这种量化方式没有考虑区域的活动特点，忽略那些可达性本身较好但是旅行分布较单一致使旅行所涉及到的热点数 a_{tj} 很少的区域。因此这会低估这些区域的可达性进而对道路的可达程度作出错误的判断。

本文则利用区域的可达率取代传统方法的可达热点数量对其可达性进行量化。区域 P_i 在特定的旅行条件 t 下于规定的旅行时间范围 TM_max 内完成的旅行的数量记为 $Acp_{_it}$，即

$$Acp_{_it}=\sum_{j=1}^{n\wedge type(t)}(trip_{java}) \tag{10}$$

$trip_{java}$ 是区域 P_i 到区域 P_j 的旅途中符合条件 t 且旅行时间小于 $TM_{_max}$ 的旅行的个数。

区域 P_i 在所有旅行条件下的综合可达率记为 $Acp_{_i}$，即

$$Acp_{_i}=\frac{\sum_{t=1}^{20}Acp_{_it}}{count_{ODi}} \tag{11}$$

$count_{ODi}$ 为由区域 p_i 出发的所有旅行的数量。

利用区域可达率量化可达性可以降低区域可达性对平均旅行时间的依赖。尽管平均旅行时间小于（大于）$TM_{_max}$，但仍有很多旅行所用时间大于（小于）$TM_{_max}$，而 $Acp_{_i}$ 是两区域间旅行时间符合时间要求的旅行占两区域间发生的所有旅行的概率，所以 $Acp_{_i}$ 较 $Ac_{_i}$ 可以更好地代表区域的可达性。除此之外考虑到热点的产生是由各区域发生的所有旅行共同聚类而成的，每个区域对热点的产生作出的贡献都不一样，即不同区域对热点的关联性是不一样的。并且每个区域的活动量及主要活动方向也不一样，因此会出现一些区域尽管大多数旅行都可以在阈值时间内到达但实际到达的热点数很少。而对于实际的需求来说不能认为这些区域可达性差。而本文提出的方法则是从旅行的角度出发，计算区域 p_i 中满足旅行时间的旅行占 p_i 所发生的所有旅行的比例，并以此量化区域可达性，这样可减少因区域旅行特点不同而产生的错误估计，更客观地代表地区在实际生活需求中的可达程度。区域可达性评估及分析

对区域可达性评估计首先要计算各区域可达率的平均值，也就是城市的平均可达率，并以此作为阈值评估区域可达性。可达率对于低于平均可达率的区域被判定为可达性较差的区域，反之可达性较好。由于本文模型的可达性评估结果是综合交通条件、旅途距离、天气情况等多种条件而得出，因此其不仅能评估出区域可达性的好坏，还可以进一步分析造成这些地区可达率高或低的原因。首先利用 OD 矩阵的旅行信息可求出区域间每一条旅行的平均旅行速度 V_k，即

$$V_k=\frac{DIS_real}{TIME_{jour}} \tag{12}$$

进而求出区域间所有旅行的平均旅行速度作为区域平均旅行速度 V_{ij}. 即

$$V_{ij}=\frac{\sum_{n=1}^{count_{ODij}}V_k}{count_{ODij}} \tag{13}$$

如果多数平均速度正常则证明是旅途距离太长导致旅行时间长，使可达率较低，如果本地区到多数地区平均速度较小则证明行车速度慢是主要影响因素。

对于区域平均速度较小的成因可通过旅行条件的对比分析得出。例如从表 1 中可以得知天气情况分为晴、雨，可以通过控制变量的方法对比同一时段两种不同天气下同一区域的可达率的变化。若两者相差不大，则可排除因天气原因导致的可达率低。若雨天可达率低，晴天可达率正常，则该区域行车速度慢的原因可能是雨天导致路况不佳，最终致使可达率降低。

3　实例研究

为了进一步说明本文所提出的区域可达性评估方法的应用过程，以沈阳市城区为例，利用沈阳市出租车 GPS 行车数据及气象数据对城市区域可达性的评估进行实例研究。实例中运用到的出租车 GPS 行车数据为沈阳市 2016 年 3 月 1 日至 2016 年 5 月 31 日的出租车 GPS 行车数据。气象数据为沈阳市同期各个气象观测站的气象数据，其中主要用到了降水量信息，以此来判断旅行条件中的晴天和雨天。

3.1　出租车 GPS 数据清洗

本实例所用到的沈阳市出租车 GPS 行车出租车 GPS 原始数据共 1925867139 条数据。但是这些数据中有一部分无效数据。因此首先要对这些原始数据进行清洗。对于原始数据的数据清洗包括删除和纠正两个步骤，每一步清洗后的数据量如图 3 所示。

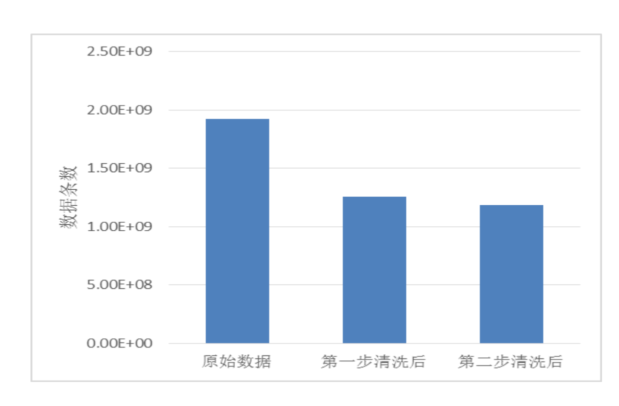


图 3　数据清洗剩余数据量

第一步数据清洗的目标是删除坐标明显错误的数据，所以过滤条件为 GPS 数据的位置坐标为明显不在沈阳范围内的坐标数据（如（0,0））。第二步清洗的目的是将行车速度和加速度明显不符合实际值的坐标点进行纠正或者清除，过滤条件 V_{max} 取值为 150km/h,加速度 M_a 取值为 5m/s。从图 3 中可以看出第一次清洗过程中删除了较多数据，经过分析发现删除的数据大多出现在夜晚，也就是大多出租车不工作的时间段，在这些时间段内数据库中的数据被错误的垃圾数据所填补。第二步纠正主要是将与正常速度与加速度差距不大的坐标点纠正并且删除差距较大的坐标点，因此其减少的数量较少，但由于 GPS 系统获取数据时就存在误差，所以经过纠正的数据量是较大的。

3.2　确定学习区域

本文将整个沈阳市按 46×46 的网格区域来划分，这样生成了 2116 个区域，在此基础上确定学习区域可以缩小学习的范围。根据 OD 矩阵中所抽取出的旅行的起始坐标与终点坐标，以 DBSCAN 方法进行热点聚类[18]，搜索半径 e 为 0.1 公里，包含的最小样本数为 15。热点的确定如图 4 所示，其中所有的黄点是聚类出的热点地区。按上述 DBSCAN 方法进行聚类所有可用的旅行信息可聚类出 4369 个热点。整个沈阳市有 521 个学习区域。



图 4　热点聚类结果

3.3　预测学习区域可达率并分析可达率低的区域

对于学习区域的可达率计算，本文首先对该区域在各旅行条件下的综合可达率进行计算。然后再计算综合可达率较差的区域进行各旅行条件下的可达率，利用这些结果可对可达率低的地区进行分析。

在区域综合可达率的计算过程中首先要确定出平均旅行时间范围的上限 TM_{max}，本文 TM_{max} 为 28_{min} 。再计算各旅行条件下的 Acp_{it}。以区域 p_i 为例，当旅行条件 t 为 WMF 时，区域 p_i 到各区域的可达率如图 5 所示。

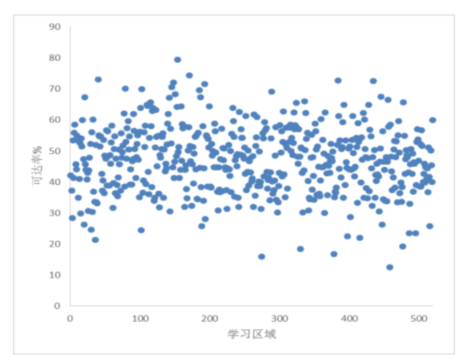
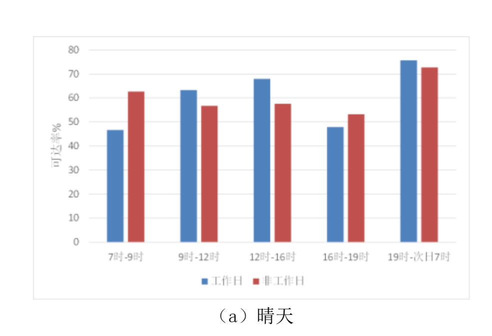


图 5　区域 p_i 到各区域的可达率示意图

各种旅行条件下区域 p_i 的平均可达率如图 6 所示。



（a）晴天

（b）雨天

图 6　各条件下区域 p_i 的平均可达率

图 6（a）为晴天工作日与非工作日各时间段的平均可达率。图 6（b）为雨天工作日与非工作日各时间段的平均可达率。由各条件下区域平均可达率可计算出该区域的综合可达率。全部 521 个区域的综合可达率如图 7 所示。

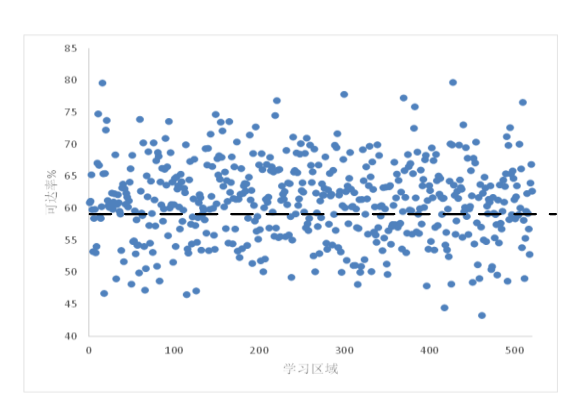


图 7　所有学习区域的综合可达率

图 7 中黑色虚线为沈阳平均综合可达率 58.6%，即这 521 个区域的综合可达率的平均值。本文以城市平均综合可达率作为区域可达性的评价标线，可达率高于标线则评价为可达性较好，反之则被评价为可达性较差。通过这样的方法可以根据量化出的可达率评价出每个区域的道路可达性，为道路交通规划提供决策依据。

从图 7 中可以看出，大部分区域的可达率高于城市平均水平，可达率较好。而对于低于平均水平的可达性较差区域可做进一步分析，首先对各旅行条件下各区域的平均速度进行研究。把 20 种出行条件分为 4 组，{WMF, WNF,WAF,WEF,WBF}、{VMF ,VNF, VAF, VEF, VBF}、{VMR, VNR, VAR, VER,VBR}、{WMR, WNR, WAR,WER,WBR}。根据式(12)(13)可求出区域间平均速度 V_{ij}，若 V_{ij} 的值并不低，则意味着区域间的旅行速度不慢，畅通性较好，那么造成可达性差的原因很可能为路途太远。如果 V_{ij} 的值较小就是说区域的可达性较差是道路不畅通所造成的，通过对比各条件下的速度可以发现是哪一时段因为什么样的原因造成的道路拥塞，利用分析出的信息对道路因地制宜的调整改造提供决策依据。

4　对比实验

为了验证本文提出的基于多元数据的区域可达性评估模型的准确率，将区域间可达率评价结果与现有应用较为普遍的基线模型和传统的轮廓测量法进行对比。对比实验数据分为训练数据和测试数据两种。将沈阳市所有出租车 2016 年 3 月-2016 年 5 月这 3 个月间 GPS 行车数据作为区间可达性评估的训练数据，将 2016 年 6 月 1 日—15 日这 15 日内的 GPS 行车数据作为测试数据。训练数据用于各种算法对区域可达性的评估，测试数据则用于来生成评估的真实值。

选取某一区域到其他 520 个区域使用基线模型与本文所提出的方法所计算的平均旅行时间计算和真实平均旅行时间进行对比，对比结果如图 8 所示。

从图 8 所对比结果可以看出红点为真实值，黑点为本文算法，蓝点为基线法。其中本文算法计算的区域间平均旅行时间与真实值相差不大，而基线法计算的时间结果总体上偏低。这是由于其平均旅行时间是基于区域间的直线距离来计算的，因此其计算出的区域平均旅行时间相对于真实值偏低。而基于基线模型区域可达性的研究十分依赖于旅行的平均时间，因此会导致其可达性评估准确性较差。而本文因考虑旅行的实际路程从而得出一个较基线模型更准确的平均旅行时间。

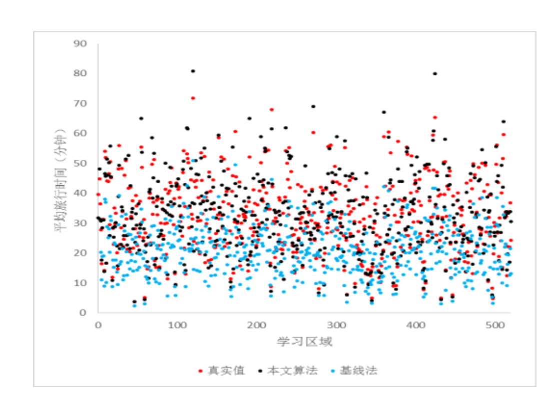


图 8　平均旅行时间对比

轮廓法采用 Ac_i 计算可达的热点个数来量化区域可达性，本文以 $Acp_{_i}$ 计算能够完成的旅行的概率并以此量化区域可达性。因为量化策略有所不同但又希望对两者的量化结果进行比较，因此在轮廓测量法计算出区域可达热点数 Ac_i 的基础上除以城市的所有热点则可以得出一个相对可达率，并以此和本文算法进行比较。本文算法与轮廓测量法计算的沈阳市 521 个学习区域综合可达率和真实值的对比如图 9 所示。

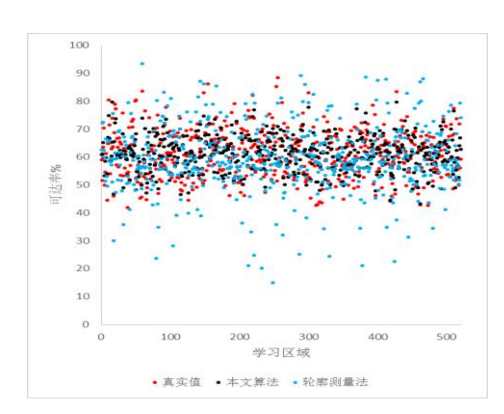


图 9　城市综合可达率对比

从图 9 中可以看出红点为真实值，黑点为本文算法，蓝点为轮廓测量法。代表本文算法的黑色点与代表真实值的红色点相差较小，因此从量化值上看本文提出的可达利率计算模型更接近真实值。而轮廓测量法是基于热点数的统计方法，容易忽略旅行的特性而导致以偏概全的情况出现，因此其可达率的分布情况较为分散。

为了验证基线法、轮廓测量法以及本文提出的模型在评估区域可达性上的差异，将三种方法判定的可达性好的区域数与真实可达性好的区域数进行对比，对比结果如表 2 所示。

表 2　不同方法的区域可达性评估结果

方法	可达性被判定为好的区域数	与真实值判定一致的区域数
真实值	320	320
基线法	396	212
轮廓测量法	293	241
本文提出的方法	323	295

由表 2 可以看出本文提出的方法评估出可达性好的区域个数最接近真实结果，并且与真实情况相吻合的区域个数也是最多的，因此本文提出的方法评估结果的准确性较好。而基线法评估出的可达性好的区域的个数是最多的，但是与实际情况相吻合的个数是最少的，因而基线法计算的准确率是最低的。这是因为低估旅行时间从而影响对区域可达性判断进而高估区域可达性造成的。轮廓测量法得到的可达性好的区域的个数较少，轮廓测量法的结果中与实际情况相吻合的个数相对较少，所以该方法的准确率也比较低。其原因是在轮廓测量法中参与判定区域可达性的可达热点数的取值只能是 0 或者涉及到的全部热点数 a_j，这样就无法避免因区域旅行分布不均等旅行特点对区域可达性判定造成的误差。

5　结束语

本文针对传统城市区域可达性评估方法不能支持多元数据分析、量化方法不够精确等问题，以出租车 GPS 行车数据、行车时段、天气等多维数据作为区域可达性的评估依据，构建了一种支持多元数据的城市区域可达性评估模型，在此基础上设计了基于多维 OD 矩阵的多元数据区域可达率计算方法，并将可达率作为区域可达性量化标准以减小旅行分布不均等旅行特点产生的误差、提高可达性量化准确性。此外，本文面向海量出租车 GPS 数据构建了一套数据清洗及有效旅行信息抽取的方法，为区域可达性的评估奠定了更有效的数据基础。经过对比实验证明相对于其他方法本文提出的基于多元数据的可达性评估模型对平均旅行时间计算的准确性提高了 18.5%~31.6%，区域综合可达率的准确率提高 12.6%~35.5%，区域综合可达性评估准确性提高 9.1%-37.8%。

由于道路交通情况复杂多样，影响交通条件的因素有很多，下一步研究还可以结合更多数据以挖掘出更多影响区域可达性的信息，如道车流量、人流量及出租车司机驾驶习惯等因素。虽然本文对量化方式加以改进，但是可达率这一量化方式并不能达到绝对的准确，在接下来的研究中可以以可达率为蓝本，优化多元数据下可达率的计算方式，以提高最终区域可达性评估的准确性。

参考文献：

[1] Coutts C J, Horner M, Chapin T. Using GIS to model the effects of green space accessibility on mortality in Florida [R]. Department of Urban & Regional Planning Faculty Publications, 2017.

[2] Cui J X, Liu F, Janssens D, et al. Detecting urban road network accessibility problems using taxi GPS data [J]. Journal of Transport Geography, 2016, 51: 147-157.

[3] Zhang A, Song S, Wang J. Sequential data cleaning: a statistical approach [C]// Proc of International Conference on Management of Data. New York: ACM Press, 2016: 909-924.

[4] Shen Q. Spatial technologies, accessibility, and the social construction of urban space [J]. Computers Environment & Urban Systems, 1998, 22 (5): 447-464.

[5] Stelder D. Regional accessibility trends in europe: road infrastructure, 1957–2012 [J]. Regional Studies, 2016 (6): 1-13.

[6] 潘竟虎，从忆波. 中国民用机场可达性与服务范围测度 [J]. 经济地理, 2015, 35 (2): 46-53.

[7] 陈艳艳，魏攀一，赖见辉，等. 基于 GIS 的区域公交可达性计算方法 [J]. 交通运输系统工程与信息, 2015, (2): 61-67.

[8] 周爱华，张景秋，张远索，等. GIS 下的北京城区应急避难场所空间布局与可达性研究 [J]. 测绘通报, 2016 (1): 111-114.

[9] Novak D C, Sullivan J L. A link-focused methodology for evaluating accessibility to emergency services [M]. [S. l.] : Elsevier Science Publishers, 2014.

[10] Qian X, Ukkusuri S V. Spatial variation of the urban taxi ridership using GPS data [J]. Applied Geography, 2015, 59: 31-42.

[11] Phiboonbanakit T, Horanont T. Who will get benefit from the new taxi fare rate?Discerning the real driving from Taxi GPS data [C]// Information and Communication Technology for Embedded Systems. 2016: 73-78.

[12] Lu Y, Li S. An empirical study of with-in day od prediction using taxi GPS data in Singapore [J]. Langmuir the ACS Journal of Surfaces & Colloids, 2014, 30 (31): 9567-9576.

[13] 徐先瑞，彭仲仁. 基于案例的城市道路行程时间预测 [J]. 交通运输系统工程与信息, 2016, 16 (4): 199-205.

[14] Mustary N R, Chander R P, Baig M N A. A performance evaluation of VANET for intelligent transportation system [J]. World Journal of Science & Technology, 2012. 2 (10) , 89–93.

[15] Laha, A. K, Putatunda. Travel time prediction for taxi-GPS data streams [J]. Iima Working Papers, 2017.

[16] Chen C, Zhang D, Zhou Z H, et al. B-planner: night bus route planning using large-scale taxi GPS traces [C]// Proc of IEEE International Conference on Pervasive Computing and Communications. 2013: 225-233.

[17] Chen C, Zhang D, Li N, et al. B-planner: planning bidirectional night bus routes using large-scale taxi GPS traces [J]. IEEE Trans on Intelligent Transportation Systems, 2014, 15 (4): 1451-1465.

[18] 陈广胜，程逸群，景维鹏. 基于 KD 树划分的云计算 DBSCAN 优化算法 [J]. 计算机工程, 2017, 43 (4): 21-27.